Vol. 20 No. 4
Jul., 2023
天 文 研 究 与 技 术
ASTRONOMICAL RESEARCH & TECHNOLOGY
第 20 卷 第 4 期
2023 年 7 月

DOI: 10.14005/j.cnki.issn1672-7673.20230323.002

deepCR 宇宙线识别方法在 CSST 巡天数据处理中的可用性及稳定性定量评价研究

林 准, 黄伟荣, 王 锋*, 邓 辉, 梅 盈
(广州大学物理与材料科学学院天体物理中心, 广东 广州 510006)

摘要: deepCR 宇宙线识别方法是哈勃空间望远镜(Hubble Space Telescope, HST)剔除宇宙线的有效方法, 但这一方法是否可以满足中国空间站望远镜(China Space Station Telescope, CSST)的要求, 始终缺乏科学的定量分析。采用哈勃望远镜的真实观测数据, 对 deepCR 宇宙线识别方法进行了深入分析, 对其稳定性和可用性进行了实测研究。结果表明, deepCR 在天空背景区域识别宇宙线的灵敏度较高, 但越靠近星像中心, deepCR 识别宇宙线的灵敏度越低。分析了宇宙线密度和测光精度的关系, 发现当宇宙线密度达到 9%时, 几乎 100%的星受到宇宙线的污染; 当宇宙线密度达到 14%时, 对于不同轮廓面积的星, 存在 20%~50%的异常测光结果。实验结果表明, deepCR 宇宙线识别模型稳定性相对较好, 一次建模后可以在较长时间内应用。但在高精度测光等应用场景仍面临着一系列问题, 需要有针对性的解决方案。

关键词: CSST; 宇宙线; deepCR; 定量评价
中图分类号: TP391.41 **文献标识码**: A **文章编号**: 1672-7673(2023)04-0333-08

宇宙线是来自宇宙的高能带电粒子[1-2], 它能穿过地面或设备留下能量痕迹。宇宙线常常干扰正常的天文观测, 尤其是在天文图像处理领域。为了提高数据的质量和可靠性, 确保科学分析的可靠性和准确性, 宇宙线必须正确地识别和剔除[3]。

最传统的做法是对同一天空区域进行多次曝光, 多张曝光图对齐并计算中值图以获得一幅无宇宙线的图像, 将每张曝光图与中值图进行比较可以识别宇宙线[4]。这类方法效果很好, 然而它并不适用于单次曝光的图像。

近年来, 人们提出了各种技术识别和剔除单次曝光天文图像中的宇宙射线。文[5]提出基于卷积的方法, 用点扩散函数减去 delta 函数以构建空间滤波器并与原图像进行卷积, 根据滤波图像的噪声特性设置阈值以识别宇宙线。文[5]指出, 程序需要进行多次迭代才能更好地识别多个像素点组成的宇宙线, 因此较为耗时, 同时该方法要求采样数据良好, 半高全宽要大于等于两个像素。文[6]根据宇宙线尖锐的边缘和无对称性两个特征, 将原始图像子采样放大并与拉普拉斯算子进行卷积, 恢复原始分辨率后得到拉普拉斯图像, 另外使用中值滤波构造精细结构, 通过设置拉普拉斯图像与噪声模型、精细结构的对比度识别宇宙线。该方法检测效果较好, 但程序需要不断迭代直到没有新的宇宙线, 对于大图程序运行时间较长, 而且在面对不同的图像数据时, 需要手动调整对比度, 以获得最好的识别效果[7-8]。文[9]考虑到宇宙线的直方图分布不是高斯分布, 提出基于图像直方图统计的方法, 该方法先将图像划分为若干个子图, 通过分析多个子图的直方图分布, 设置阈值将偏离分布的像素点识别为宇宙线。文[9]指出, 该方法适合处理光谱图像数据, 而且运行速度快, 但对于点扩散函数较窄的图像, 宇宙线识别效果不如文[5]和文[6]。

基金项目: 国家自然科学天文联合基金 (U1831204, U1931141); 国家自然科学基金国际合作项目 (11961141001) 资助.
收稿日期: 2023-03-02; 修订日期: 2023-03-15
作者简介: 林 准, 男, 硕士研究生, 主要研究天文技术与方法. Email: linzhun@cnlab.net.
* 通信作者: 王 锋, 男, 教授, 主要研究天文技术与方法. Email: fengwang@gzhu.edu.cn.

随着深度学习技术的发展，人们开始利用深度学习方法识别宇宙线。其中，deepCR[10]由于其高精确度和高效率而受到极大关注。整个框架包括两个独立的深度神经网络，分别用于标记宇宙线和标记过后的图像恢复。deepCR已在哈勃空间望远镜高级巡天相机(Advanced Camera for Surveys，ACS)/广域通道(Wide Field Channel，WFC)(F606W滤波器)的数据上展示了比拉普拉斯边缘检测算法更高的召回率以及更快的处理速度。

中国空间站望远镜[11]是我国载人航天工程规划建设的重大科学项目，主要任务是进行大规模天文巡天。然而，在巡天观测中，宇宙线干扰是一个严峻的问题。中国空间站望远镜主巡天相机由30块探测器拼接组成，覆盖区域大约为1.1×1.2平方度，焦面感光面积约为234 000 mm^2。根据与中国空间站望远镜有相近轨道高度的哈勃空间望远镜的数据估计，每块探测器在150 s的曝光过程中，有超过20万个像元受到宇宙线的影响。宇宙线会破坏这些像元的读数，因此需要对这些像元的位置进行标定，以避免对科学数据的测量产生影响。在巡天模式中，由于每个探测器只对相应天区覆盖一次，无法使用多次曝光合并这种传统的去除方式，只能利用单次曝光图像实现宇宙线去除。

尽管deepCR取得了较好的结果，在实际应用中它的真实效果仍缺少定量分析，这个方法是否可以应用于未来的中国空间站望远镜数据处理值得研究。本文基于哈勃空间望远镜的观测数据，针对deepCR进行深入分析，对其实际应用的稳定性和可用性进行了系统的定量研究。

1 deepCR宇宙线识别方法

deepCR宇宙线识别模型由两个独立的深度神经网络组成，分别是deepCR-mask和deepCR-inpaint[10]。deepCR-mask对输入图像预测每一个像素被宇宙线影响的概率，而后用0.5倍阈值将概率图转换为二值图，0代表非宇宙线，1代表宇宙线。deepCR-inpaint预测输入图像中标记为宇宙线的像素点在没有宇宙线干扰时的值，以此恢复图像在没有宇宙线干扰时的情况。这两个网络都基于UNet结构搭建。

deepCR-mask模型的训练数据包括受宇宙线影响的图像和与之对应的宇宙线二值标签图。选择多次曝光数据作为训练数据，通过比较每张曝光图与中值图的方法制作准确的宇宙线二值标记。我们直接使用文[10]提供的基于2005年的HUBBLE ACS/WFC(F606W)数据训练好的模型进行后续测试。后续测试的数据在第2节介绍。

2 测试数据准备

为了与文[10]保持一致，我们使用HUBBLE ACS/WFC(F606W)的观测数据，随机选取20组，每组由视场相同并且经过校准的6次曝光的科学图像构成(见图1)。图1为每组数据的观测日期以及平均曝光时间。AstroDrizzle[12]数据管线可以自动对多张观测图像进行天空投影对齐，随后进行中值合并，得到没有宇宙线的中值图，再将中值图映射至每张原始的观测图像，得到多张映射中值图。使用5倍均方根(Root Mean Square，RMS)和1.5倍均方根作为第1道和第2道阈值，比较映射中值图与原始观测图像的差异，从而标记宇宙线的位置。用映射中值图上的像素值替换原始观测图像中的宇宙线，最终得到干净图像。此外，我们还将哈勃望远镜观测数据(flc. fits文件)中的数据质量数组作为坏像素的掩码，并为大于70 000e^-的像素创建7×7的饱和掩码，以保证这些异常像素不参与后续的模型评价。

基于AstroDrizzle数据管线，我们得到了每张原始观测图像剔除宇宙线后的干净图像，把这些干净图像作为基准图像，以基准图像上的测光结果作为基准的测光结果，开展后续的定量评价实验。

3 deepCR定量评价

3.1 基于深度学习指标的定量评价

由于宇宙线在全图的占比小，导致正负样本不均衡，因此精确率和召回率是最重要的衡量指标。

精确率 = $TP/(TP+FP)$，召回率 = $TP/(TP+FN)$，其中，TP 代表正确识别为宇宙线的宇宙线像素的数量；FP 代表错误识别为宇宙线的非宇宙线像素的数量；FN 代表错误识别为非宇宙线的宇宙线像素的数量。

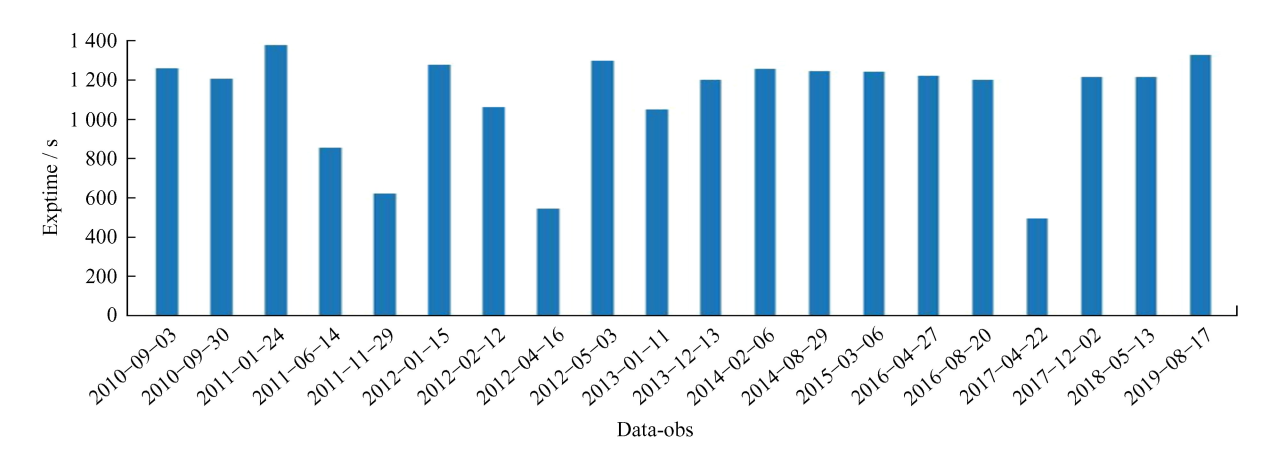


图 1　实验数据描述

Fig. 1　Experimental Data Description

利用文[10]的模型，我们首先对所有测试数据进行全图像区域的宇宙线识别评估。结果显示召回率达到 88.8%，精确率为 95%。这表明模型不仅可以较准确地分辨宇宙线，而且具有较高的识别灵敏度。然而，由于宇宙线大部分落在天光背景上，可能导致识别效果被高估。为了得到更可信的精确率和召回率，我们进一步对星的区域进行评估。使用 SEP[13-14] 计算干净图像背景的均方根，并使用均方根的 3 倍、5 倍、10 倍等阈值来提取星的连通区，倍数越大，表示提取的连通区域越接近星的中心。我们用这些连通区作为掩模，得到位于星上不同区域的宇宙线的识别情况。

表 1 展示了模型对位于星上不同区域的宇宙线的识别性能。我们发现，与整个图像区域的宇宙线识别效果相比，deepCR 对落在星上的宇宙线的识别灵敏度与准确率显著下降(越接近星像中心，召回率与精确率越低)。说明 deepCR 方法在星的中心区域很容易漏掉宇宙线。

表 1　deepCR 模型表现随距星像中心的变化

Table 1　Variation of deepCR performance with distance from the centre of stars

Cosmic ray position	Full graph	3 RMS	5 RMS	10 RMS	20 RMS	40 RMS	80 RMS	160 RMS
Recall/(%)	88.8	85.5	83.5	79.8	74.2	66.9	59.7	56.2
Precision/(%)	95.0	87.6	87.7	88.3	86.4	81.1	74.6	72.8

3.2　基于测光的定量评价

虽然精确率和召回率是评估模型效果的重要指标，但它们并不能反映模型在实际应用场景的效果。因此，我们基于测光结果进一步分析了 deepCR 的宇宙线剔除效果。

图 2 展示了测光的过程。本文 SEP 提取源的标准为(1)单像素高于 3 倍背景噪声；(2)连通区域像素数大于 16；(3)只对点源进行分析，每个源的长轴与短轴之比为 0.8~1.2；(4)为了避免图像边沿不可靠的测光结果，提取范围为距离 CCD 四条边大于 128 个像素的位置。另外，对同一组数据中每张图的星进行匹配，把坐标距离小于 0.1″的星判定为同一颗星。为了得到更准确的宇宙线剔除效果，同一组数据中的每颗相同的星均采用相同的测光中心坐标和测光半径。

我们定义原始图流量与干净图流量的差值大于 3 倍流量噪声的星为受宇宙线污染的星，图 3 为测试数据的测光结果分析。(a)图中，黑线为原始图像的测光结果和干净图像中的测光结果的比值；

(b)图中，蓝线为 deepCR 对原始图像剔除宇宙线后的测光结果和干净图像的测光结果的比值(下文简称为“流量比值”)。如果流量比值越接近 1，说明 deepCR 的效果越接近基准结果，比值小于 1，说明 deepCR 把星信号误判为宇宙线，比值大于 1 则说明 deepCR 剔除宇宙线不干净。

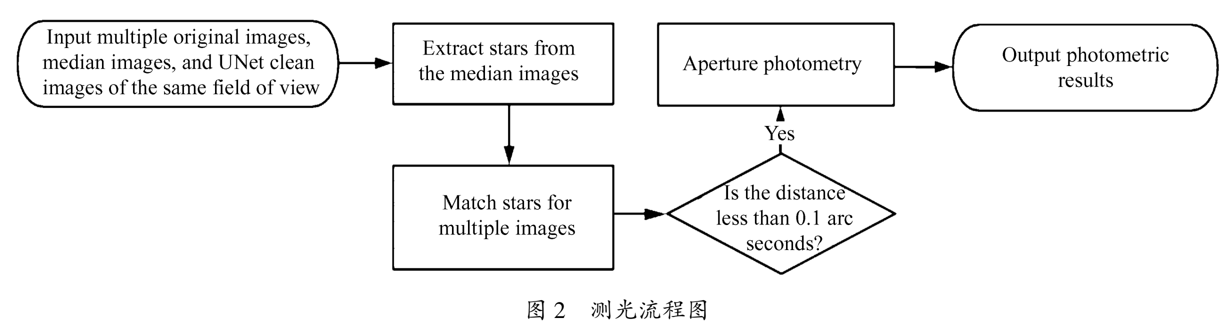


图 2 测光流程图

Fig. 2 Flowcharts of photometry

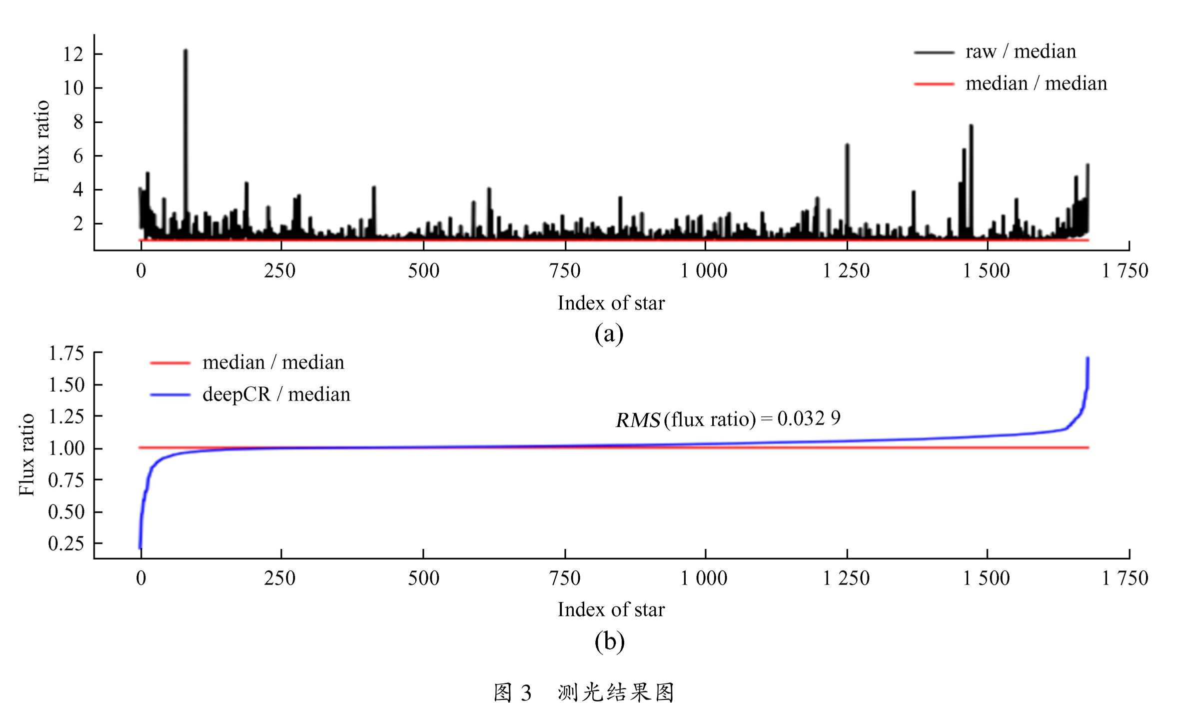


图 3 测光结果图

Fig. 3 Result of photometry

经过统计发现，每组数据的异常星占所有星的比例在 13%与 65%之间，可见大部分星受到宇宙线的污染。在这些被宇宙线影响的异常星中，有 10%的星 deepCR 剔除宇宙线后的测光结果与基准测光结果比值大于 1.1 或小于 0.9；有 55%的星 deepCR 剔除宇宙线后的测光结果与基准测光结果比值在 0.97 与 1.03 之间。尽管整体的剔除效果较好，但仍有部分星被错误剔除或者没有剔除宇宙线，从而导致测光结果异常。

进一步，我们检查了流量比值明显小于 1、明显大于 1 的星的图像(见图 4、图 5)。在每一行的图像中，第 1 个子图为原始图像；第 3 个子图为干净图像；第 2 个子图为第 1 个子图减去第 3 个子图的差，该差值表示真实宇宙线的位置；第 5 个子图是 deepCR 剔除宇宙线后的图像；第 4 个子图为第 3 个子图与第 5 个子图的差(取绝对值)，该差值表示 deepCR 误识别的宇宙线或者漏识别的宇宙线。在流量比值明显小于 1 的例子中发现，deepCR 可能错误地把轮廓较小的整颗星识别为宇宙线(参考图 4)，这极大影响了测光精度。而当宇宙线落在星的中心区域时，deepCR 往往没有成功识别出该宇宙线(参考图 5)。在另一些情况下，当宇宙线落在星的边缘时，deepCR 的识别效果良好；落在星上的宇宙线较少或较弱时，无论 deepCR 识别宇宙线的效果如何，都不会对测光结果造成明显影响。

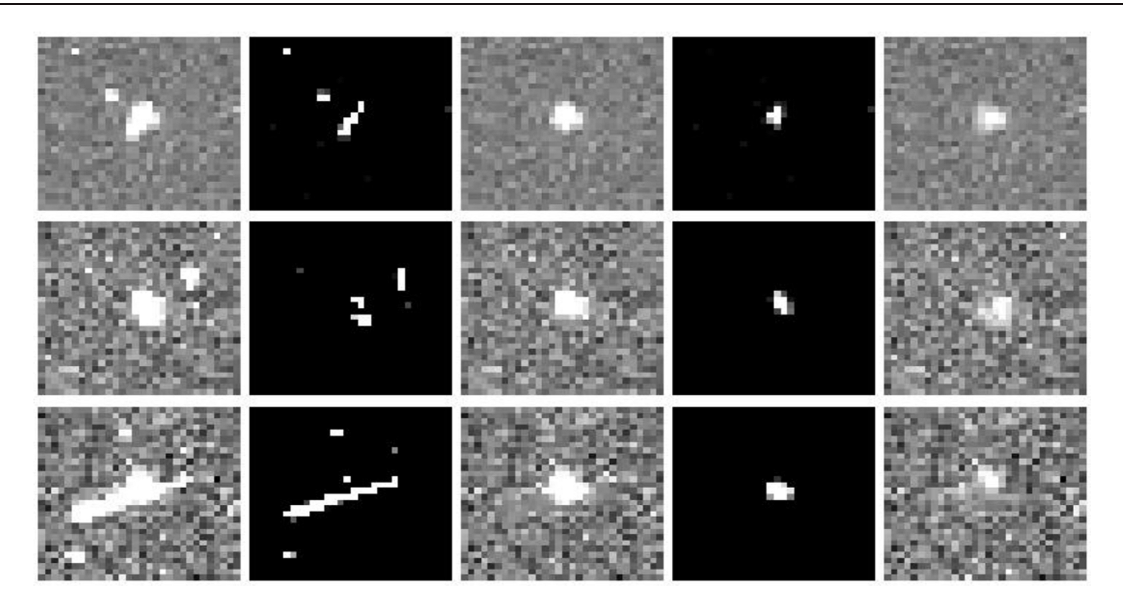


图 4　流量比值明显小于 1 的典型情况

Fig. 4　Diagram of stars with a flux ratio significantly less than 1

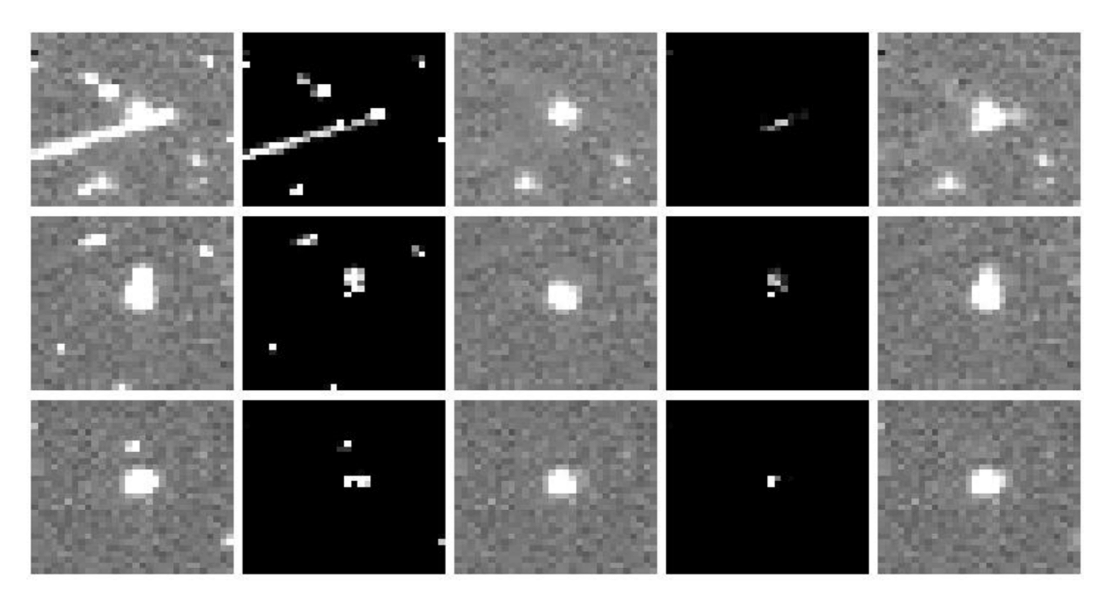


图 5　流量比值明显大于 1 的典型情况

Fig. 5　Diagram of stars with flux ratio significantly greater than 1

以上的分析说明，落在星上的宇宙线，模型识别效果并不好，因此，宇宙线的密度是测光精度的重要影响因素。图 6 展示了宇宙线密度对测光精度的影响，其中横轴代表宇宙线全图占比，宇宙线占比 4%时对应的曝光时间约为 1 200~1 500 s；纵坐标分别代表被宇宙线污染的星的比例(星被宇宙线污染的判别标准为原始图像测光结果在干净图像测光结果的 3 倍流量误差的范围外)和测光结果异常的比例(异常的判别标准为 deepCR 测光结果在干净图像测光结果的 3 倍流量误差的范围外)。(a)图中，红线为线性拟合结果，皮尔逊相关系数 $r=0.86$，$p=7\times10^{-16}$；(b)图中，黑线、红线和蓝线分别为 16~50，50~200 和 200~400 个像素大小的星的线性拟合结果，其皮尔逊相关系数 r 和 p 值分别为 $(0.80, 2\times10^{-6})$，$(0.95, 2\times10^{-12})$ 和 $(0.88, 1\times10^{-8})$。从(a)图可以看到，当宇宙线全图占比为 9%时，几乎 100%的星被宇宙线污染了。从(b)图可以看到，当星的轮廓面积为 200~400 个像素且宇宙线全图占比约 14%时，大约 50%的星测光结果异常。由于宇宙线的密度随着曝光时间的增加而增大，因此，中国空间站望远镜巡天过程中不应该采用过长的曝光时间。

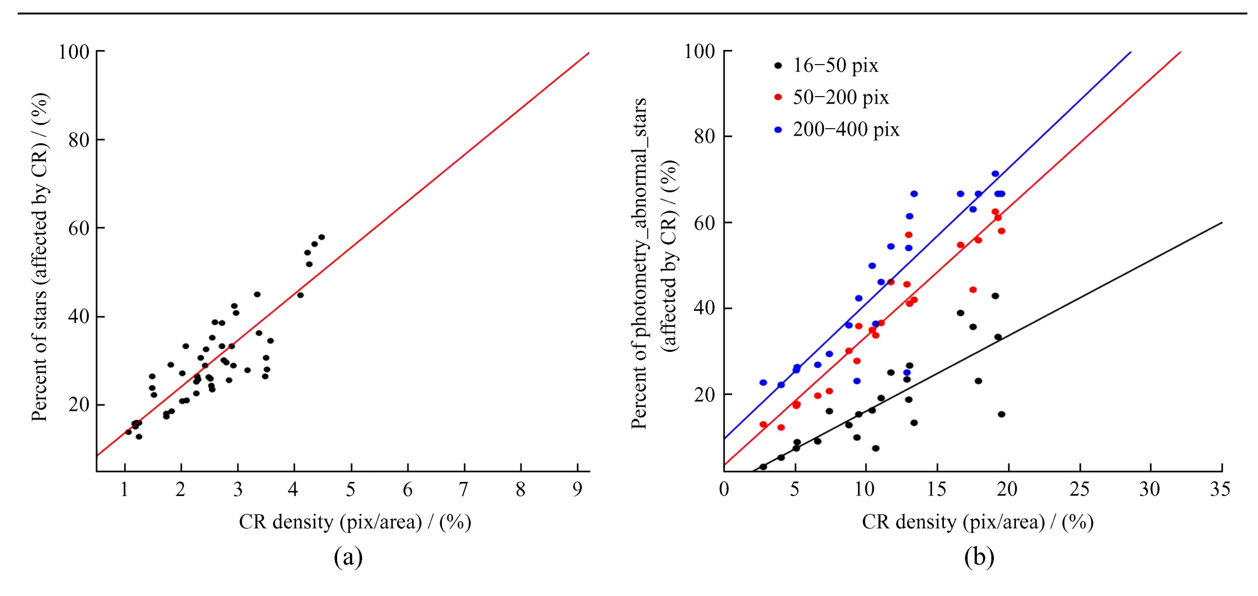


图 6 不同宇宙线密度时的测光结果异常的比例，不同颜色的点代表不同大小的星

Fig. 6 Proportion of anomalies in photometric results at different cosmic ray densities, different colored dots represent stars of different sizes

3.3 deepCR 模型稳定性分析

天文图像的质量和稳定性往往难以保证，因此一个好的深度学习模型应该具有高度的稳定性，先前缺乏对模型稳定性的研究。因此，本文通过基于深度学习评估指标和测光精度两个方面分析模型的稳定性。

图 7 展示了 2010 年 9 月至 2019 年 8 月期间 20 组观测数据基于星上的宇宙线统计的 F_1 分数和 F_2 分数。F_1 的值为 $(85 \pm 2.4)\%$，F_2 的值为 $(88 \pm 2.1)\%$。可以看到随着时间变化，F_1 分数和 F_2 分数没有明显的变化。

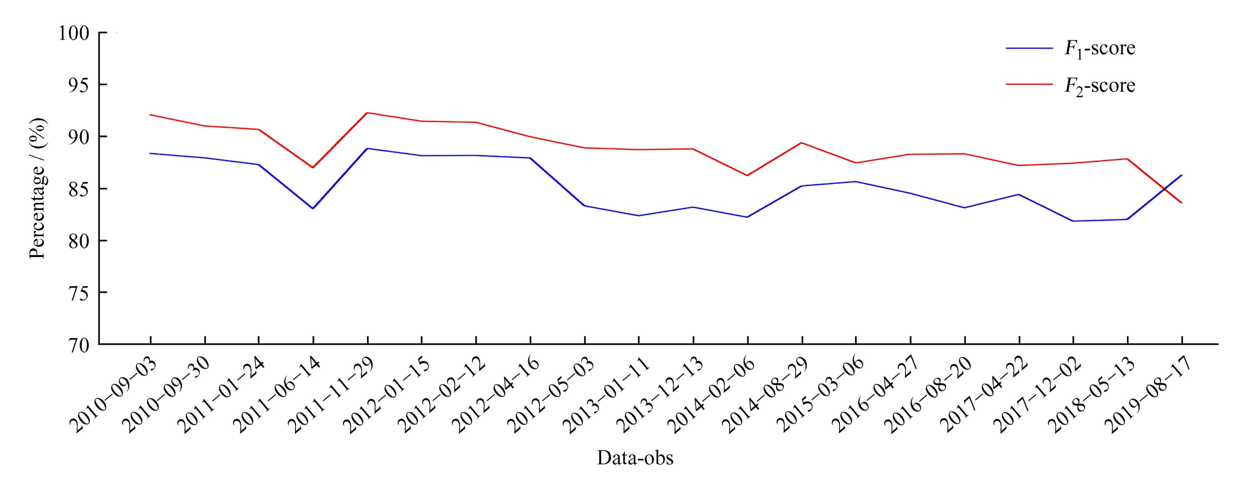


图 7 F_1 分数和 F_2 分数随观测日期的变化

Fig. 7 F_1-score and F_2-score of data from different observation times

图 8 中，横轴的日期跨度从 2010 年 9 月到 2019 年 8 月，纵坐标是每一组数据中所有星的流量比值的均方根(参考图 3 的做法)，蓝点代表测光精度；红线是蓝点数据的均值；绿线代表均值 ±3 倍标准差。另外，为了得到更加准确的均方根，去除数据中最大和最小的 5% 的数据再进行均方根的计算。结果表明，每组数据的均方根基本在误差范围内，即 deepCR 剔除宇宙线的测光精度和探测器的工作年龄没有显著关系。

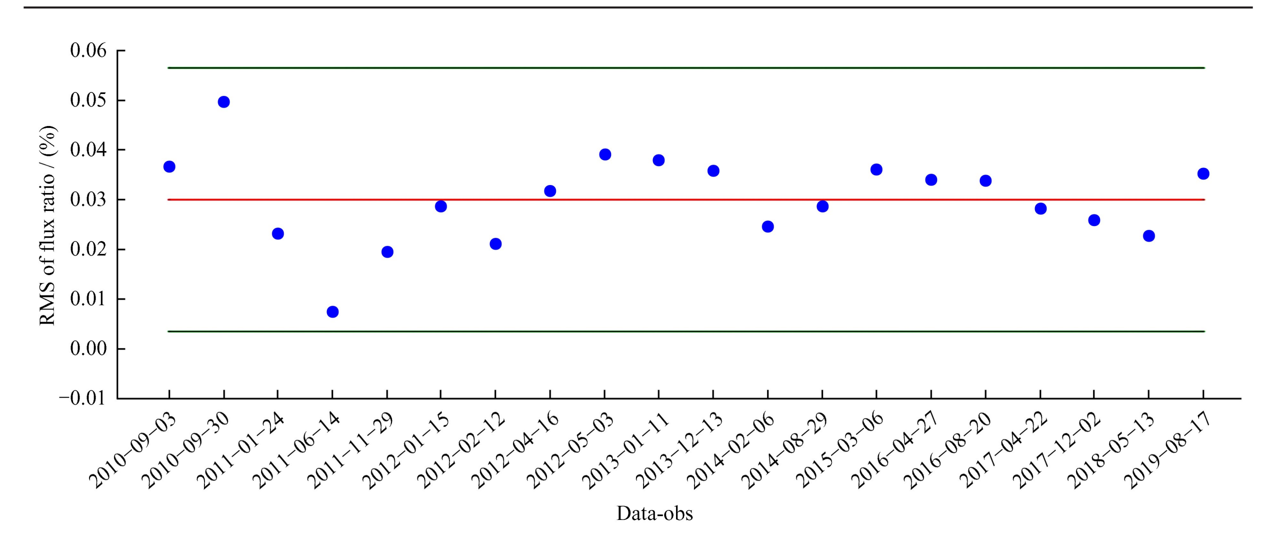


图 8　不同观测时间数据的测光精度(RMS)图

Fig. 8　Photometric accuracy (RMS) of data from different observation times

4　总　结

本文系统定量地分析了 deepCR 方法对宇宙线识别的结果。2010 年至 2019 年期间哈勃空间望远镜观测数据的 F_1分数、F_2分数和测光精度均没有明显的变化。总体来看，deepCR 方法有较好的模型稳定性。实验表明，越接近星像中心，deepCR 识别宇宙线的灵敏度越低，与大于 3 倍背景均方根的星上区域相比，在大于 160 倍背景均方根的星上区域，deepCR 识别宇宙线的灵敏度从 88.8%下降到 56.2%。此外，本文研究了宇宙线密度和受到宇宙线污染的星的比例、deepCR 方法导致异常测光结果的比例的关系。研究发现，当宇宙线像素的全图占比为 9%时，几乎 100%的星受到宇宙线污染；当宇宙线占比为 14%时，在轮廓面积为 200~400 个像素的星中，大约有 50%的星测光结果异常；在轮廓面积为 16~50 个像素的星中，大约有 20%的星测光结果异常。由于图像中的宇宙线占比越大，星像上宇宙线的识别效果越差，因此，建议中国空间站望远镜在巡天过程中尽量采用较短的曝光时间，以控制图像中宇宙线的数量。

参考文献：

[1]　卢伯强. 宇宙线和暗物质探测研究 [D]. 南京：南京大学, 2017.
LU B Q. Study of cosmic rays and dark matter detection [D]. Nanjing: Nanjing University, 2017.

[2]　MILES N D, DEUSTUA S E, TANCREDI G, et al. Using cosmic rays detected by hst as geophysical markers. I. detection and characterization of cosmic rays [J]. The Astrophysical Journal, 2021, 918(2): 86-113.

[3]　冯海霞, 陈建军, 邓建榕, 等. CCD 图像中宇宙线 μ 子甄选技术 [J]. 天文研究与技术, 2020, 17(2): 201-209.
FENG H X, CHEN J J, DENG J R, et al. Cosmic-ray muons extraction technology in CCD image [J]. Astronomical Research & Technology, 2020, 17(2): 201-209.

[4]　WINDHORST R A, FRANKLIN B E, NEUSCHAEFER L W. Removing cosmic-ray hits from multi-orbit hst wide field camera images [J]. Publications of the Astronomical Society of the Pacific, 1994, 106(701): 798-806.

[5]　RHOADS J E. Cosmic-ray rejection by linear filtering of single images [J]. Publications of the Astronomical Society of the Pacific, 2000, 112(771): 703-710.

[6] VAN DOKKUM P G. Cosmic-ray rejection by laplacian edge detection [J]. Publications of the Astronomical Society of the Pacific, 2001, 113(789): 1420-1429.

[7] 刘婷婷, 彭青玉. 消除 CCD 图像中宇宙射线的算法的比较 [J]. 天文研究与技术, 2010, 7(2): 140-149.
LIU T T, PENG Q Y. Comparison of CCD image cosmic-ray removal algorithms [J]. Astronomical Research & Technology, 2010, 7(2): 140-149.

[8] FARAGE C L, PIMBBLET K A. Evaluation of cosmic ray rejection algorithms on single-shot exposures [J]. Publications of the Astronomical Society of Australia, 2005, 22(3): 249-256.

[9] PYCH W. A fast algorithm for cosmic-ray removal from single images [J]. Publications of the Astronomical Society of the Pacific, 2003, 116(816): 148-153.

[10] ZHANG K M, BLOOM J S. Deepcr: cosmic ray rejection with deep learning [J]. The Astrophysical Journal, 2020, 889(1): 24-32.

[11] 詹虎. 载人航天工程巡天空间望远镜大视场多色成像与无缝光谱巡天 [J]. 科学通报, 2021, 66(11): 1290-1298.
ZHAN H. The wide-field multiband imaging and slitless spectroscopy survey to be carried out by the Survey Space Telescope of China Manned Space Program [J]. Chinese Science Bulletin, 2021, 66(11): 1290-1298.

[12] HACK W J, DENCHEVA N, FRUCHTER A S, et al. AstroDrizzle: more than a new MultiDrizzle [C] // Proceedings of the American Astronomical Society Meeting. 2012.

[13] BARBARY K. SEP: source extractor as a library [J]. Journal of Open Source Software, 2016, 1(6): 58.

[14] BERTIN E, ARNOUTS S. SExtractor: software for source extraction [J]. Astronomy and Astrophysics Supplement Series, 1996, 117(2): 393-404.

A Study on Quantitative Assessment of Usability and Stability for deepCR Cosmic Ray Identification Methods in CSST Survey Data Processing

Lin Zhun, Huang Weirong, Wang Feng*, Deng Hui, Mei Ying

(Center for Astrophysics, School of Physics and Materials Science, Guangzhou University, Guangzhou 510006, China, Email: fengwang@gzhu.edu.cn)

Abstract: The deepCR cosmic ray identification method could effectively remove cosmic rays from the Hubble Space Telescope (HST). However, there needs to be more quantitative analysis on whether this method can meet the China Space Telescope (CSST) requirements. In this paper, we analyze the deepCR cosmic ray method in-depth using real observational data from the Hubble Telescope. We conducted an empirical study of its stability and usability. The results show that deepCR performs well in identifying the cosmic ray in the sky background region, but its sensitivity decreases as it approaches to the centre of star. We analyze the correlation between the cosmic ray density and the photometric accuracy, demonstrating that when the cosmic ray density reaches 9%, alomost all stars (100%) are contaminated by cosmic rays; when the cosmic ray density reaches 14%, there are abnormal photometric results (20%-50%) for stars with different contour areas. The analysis shows that the deepCR cosmic ray identification model is relatively stable and can be applied over a longer period of time after one model. However, it still faces a series of problems in application scenarios, such as high-precision photometry, which need to be addressed in the near future.

Key words: CSST; cosmic ray; deepCR; quantitative assessment